Lab: Binomial Proportion CI

1. Free Throws

We have 2023-2024 free throw data

Raw Data:
- Row = Player-team
- G = games played
- FT = free throws made per game
- FTA = free throws attempted per game

- Create the dataset:
  - Row = player
  - FT = total # made free throws
  - FTA = total # attempted free throws
  - $\hat{p}$ = FTP = seasonal FT percentage

- Plot: $\hat{p}$ (x axis) versus player name (y axis) and overlay the Wald C.I.s and Agresti CI's.

Filter out all players below a certain threshold of free throw attempts. Thoughts?

2. Simulation Study

Discretize the interval $[0,1]$ into small bins. For each p in $\{0, \text{midpoint of each bin}, 1\}$ and each n in $\{10, 100, 1000, 10000\}$ and maybe some others, generate n free throws (n Bernoulli(p) coin flips) $\{X_i\}_{i=1}^n$. Compute the Wald CI and Agresti CI from the simulated data. For each (n,p) combo, repeat this $M=100$ times.

For each (n,p), in what percentage of simulations does the true p lie in each CI (i.e., estimate the coverage)?

Plot coverage (y axis) versus p (x axis) for each n and interval method.

3. Math HW

- Prove that for $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$ $\hat{p} = \overline{X}$ is the MLE (maximum likelihood estimate) of p; it maximizes the probability of observing the data given that parameter,

$$\hat{p}_{MLE} := \underset{p \in [0,1]}{\text{argmax}} \ \mathbb{P}(X_1, \ldots, X_n \mid p)$$

- After how many sports bets can you be confident that you're actually good at sports betting? Assume you only bet on -110 game winner outcomes.

Hint: Model the Return on a \$1 bet by $R = \begin{cases} -1 & \text{if lose, w.p. } (1-p) \\ +\frac{100}{110} & \text{if win, w.p. } p \end{cases}$

To break even, need $\mathbb{E}R = \frac{100}{110}p - (1-p) = 0 \implies p = \frac{110}{210}$.

- Basketball win probability via Normal approximation:

The points scored by a team in a basketball game is the sum of the points scored in each possession. If you assume this is iid, by CLT the points scored through n possessions is approximately normal. Devise a formula for the probability team 1 beats team 2 assuming each team's score is independent, n possessions remain in the game, S_1, S_2 are the points scored of teams 1 and 2, μ_1, μ_2 are the mean pts scored in a possession by teams 1, 2 and σ_1, σ_2 are the s.d.'s of poss pts scored. Test it for sensible values for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, S_1, S_2, n$ does it seem reasonable? When/why is it good and bad?

- Read the slides below entitled "The Normal Distribution in Sports & Z-Scores"

The Normal Distribution in Sports & Z-Scores

THE EMPIRICAL RULE:

A Connection between Quantiles, the Mean and the Standard Deviation

particularly sums of iid RV's, but often stuff involving humans (e.g. heights, talent, sports skill, etc)

For many datasets (the vast majority but not all) there is a simple connection between approximate percentiles and the mean and SD:

1. The majority of your data (about 2/3) is within 1 SD of the mean.

2. Most of your data (about 95%) is within 2 SD of the mean.

3. Almost none of your data (just a few per 1000) is extreme- more than 3 SDs from the mean.

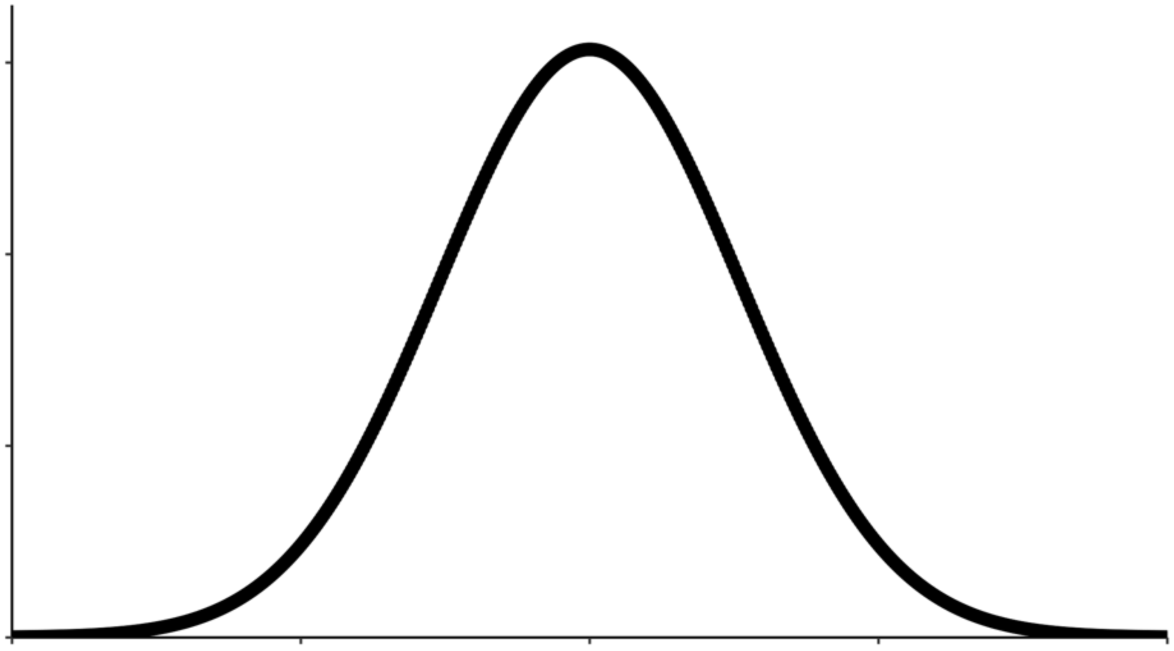
The Empirical rule is a descriptive tool; it is a way to describe a data set with just two numbers. It is remarkably useful, as we shall see through examples:

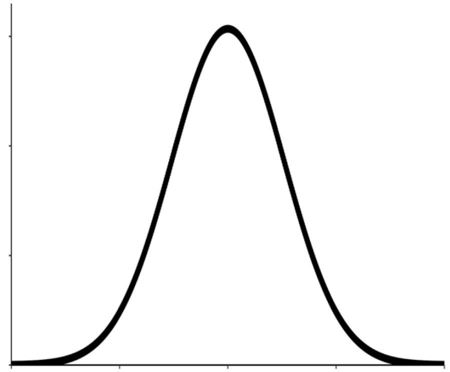
- The first rule describes where the data is mainly- within 1 SD of the mean.
- The second rule describes where the data is mostly, within 2 SDs of the mean.
- The third rule describes where the data is almost always not: more than 3 SDs from the mean.

The empirical rule makes it possible to use two numbers to know what is typical, unusual and exceedingly rare in the data.

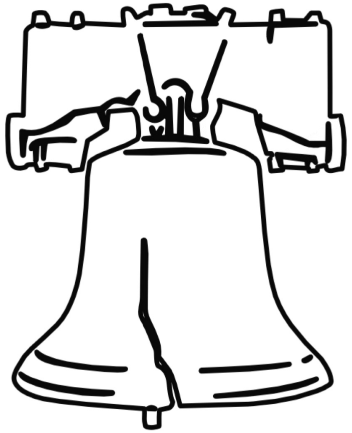
The Bell Curve

This shape often approximates the shape of histograms of many data sets that occur naturally. They are also called **Normal Curves**.





Bell Shaped Curve



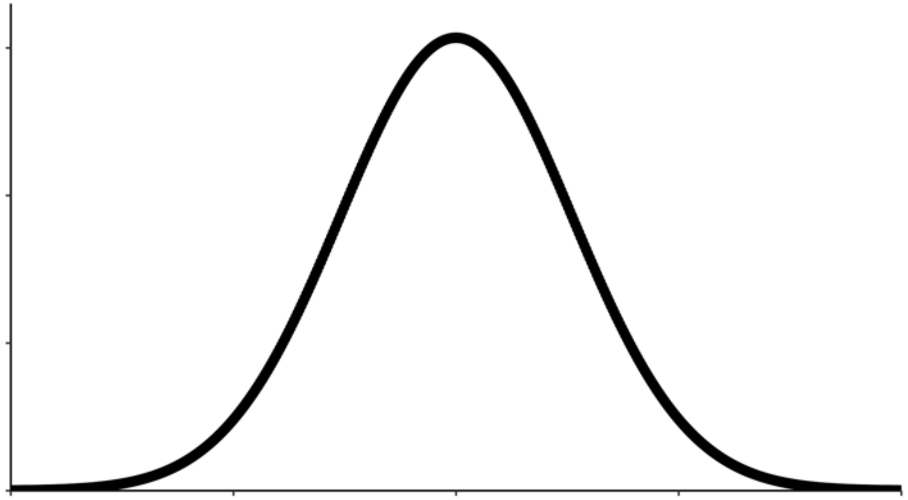
Liberty Bell in Philadelphia.

The closer the histogram for the data is to the Bell-shaped curves, the better the empirical rule is as an approximation.

The Bell Curve

The "Bell-Curve" or "Normal" curve can be scaled and shifted, but its basic shape is called the "Standard Normal Curve" and it has a mathematical equation that defines it:

$$\varphi(\chi) = \frac{e^{-\chi^2}}{\sqrt{\pi}}$$



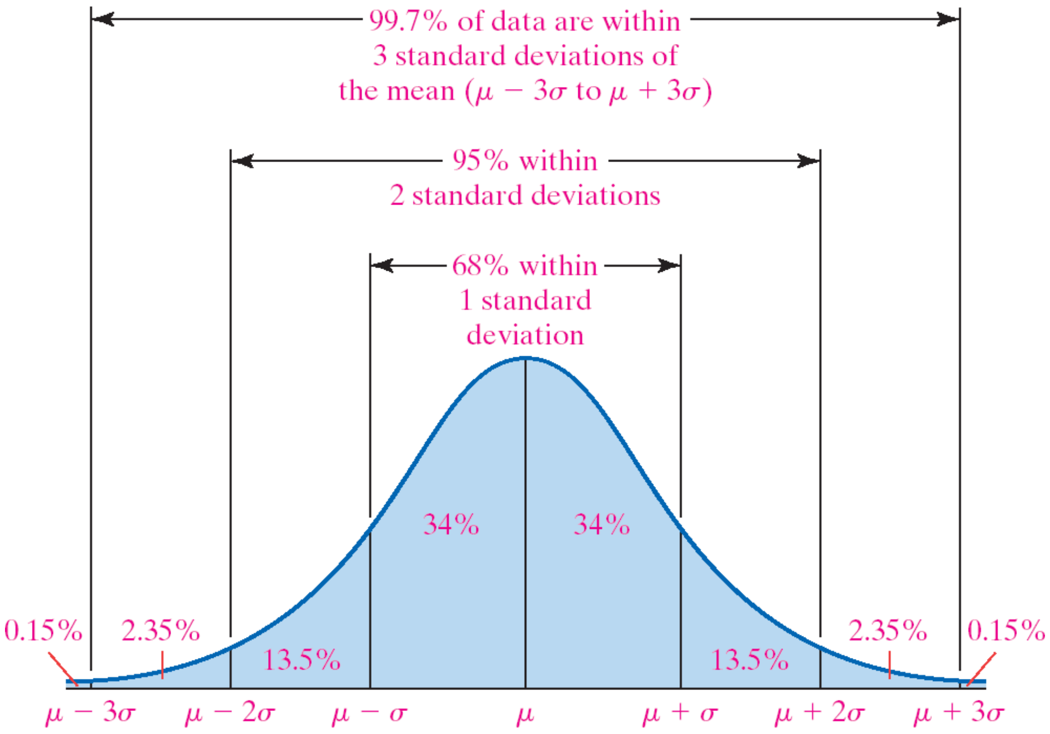
Graph of Standard Normal Curve

The standard Normal curve is centered at 0 and the total area under the curve is 1.0. The area between any two points cannot be computed analytically (there is no formula) but it can be computed numerically.

The Bell Curve

1. Area under the curve between [-1, +1] SD is .682 (68.2% of the total) **(majority)**
2. Area between [-2, +2] is .954 (95.4% of total area) **(most)**
3. Area between [-3, +3] is .997 (99.7% of total area) **(almost all)**

The Normal curve can be centered at any value: usually denoted with the Greek letter μ.
It can be scaled by any value, denoted with the Greek letter σ.



Standard Units and Z-scores

The empirical rule can be applied to any data point by counting how many standard deviations it is from the mean.

For example, the 2001 Seattle Mariners had a winning percentage of 71.6% which is 3.04 standard deviations above the mean.

This process, which changes the units of the data to a SD scale, is called **standardization**.

Mathematically, standardization is the transformation of any data point x into "standard units" z by subtracting the mean and dividing by the SD's:

$$Z = \frac{x - \bar{x}}{s}$$

Case Study: Beane vs. Cashman
Excess Wins/Season After Adjusting for Payroll

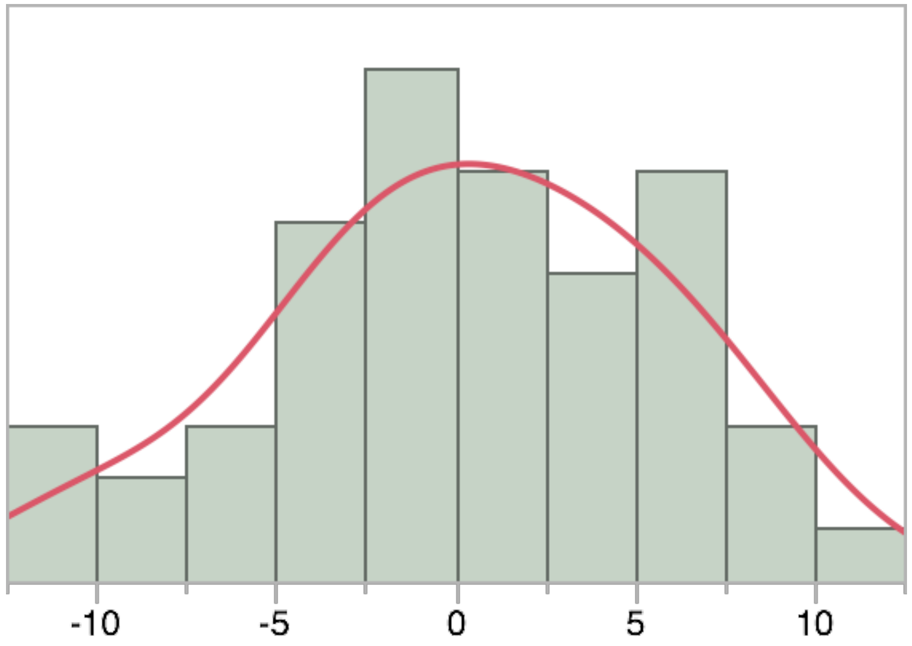
Adjusting the Data: a huge idea.

It is crucial to adjust the data so that we can standardize and account for confounding factors to find our true answer. So we compute the expected number of wins that a team *should* have given the size of their payroll; the higher the payroll, the more wins a team *should* have.

Then, once this is found, we can figure out how a team differed from this number: did they have more wins than they should? Less?

Case Study: Beane vs. Cashman, Excess Wins/Season After Adjusting for Payroll

Mean = 0.319
SD = 5.36



100%	Maximum	10.44
99.5%		10.44
97.5%		9.98
90%		7.21
75%	Quartile	4.48
50%	Median	0.109
25%	Quartile	-3.38
10%		-7.72
2.5%		-11.35
0.5%		-11.65
0	Minimum	-11.65

Brian Cashman: 5.4 extra wins/season
Billy Beane: 10.4 extra wins/season

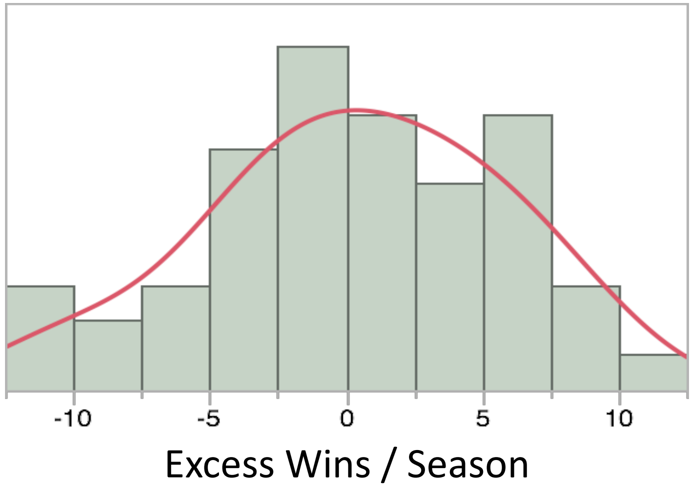
Case Study: Beane vs. Cashman, Excess Wins/Season After Adjusting for Payroll

Adjusting the Data: a huge idea.

Obviously, the Yankees *should* have the most wins given their payroll, but what if they underperformed those expectations? Then their excess wins would be negative- the team is doing worse than it should, given its payroll.

If it's positive, then the team is outperforming its payroll.

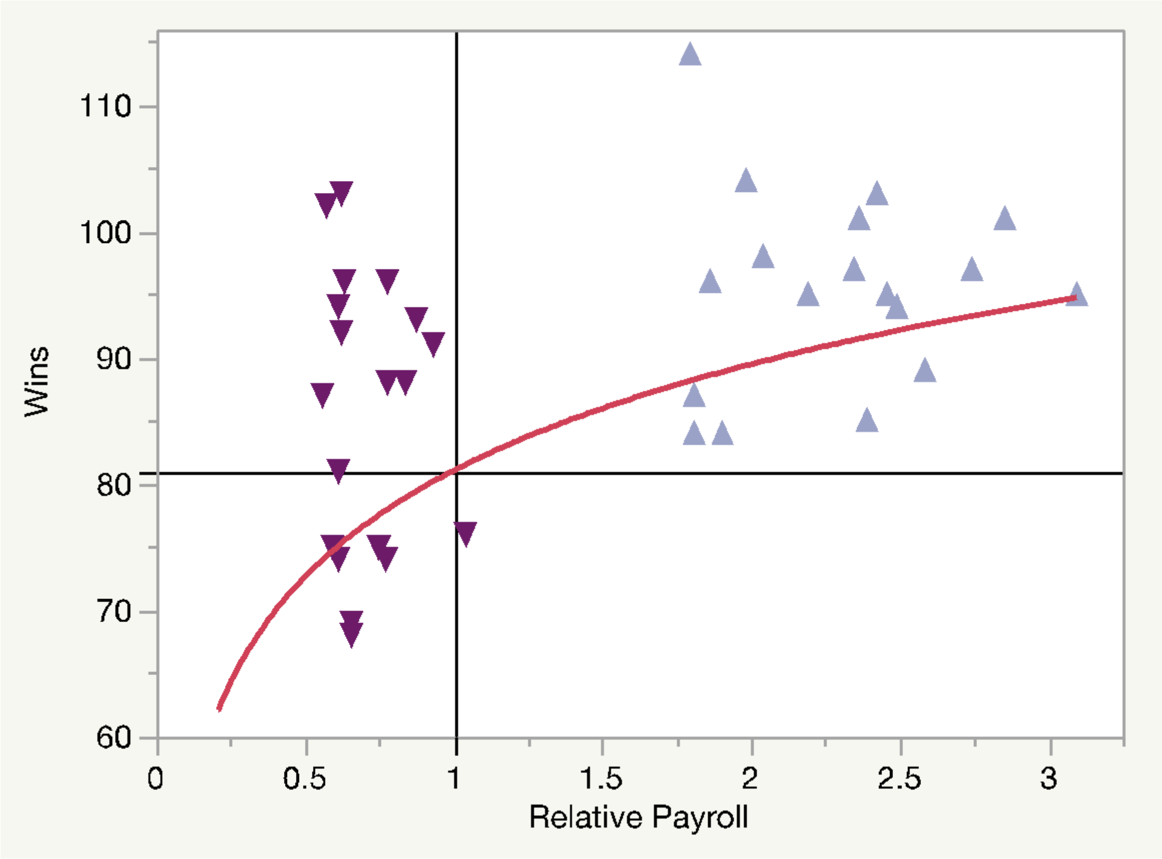
This histogram reveals the distribution of **excess wins**; as you can see, the median is almost 0- about half the teams outperform expectations and about half underperform.



Case Study: Beane vs. Cashman, Excess Wins/Season After Adjusting for Payroll

Adjusting the Data: a huge idea.

Now we can more easily compare the A's and the Yankees, because we can compare how well each team actually did to how well each team *should* have done given the payroll.



The red curve is the expected number of wins earned at a given relative payroll.

Standard Units: The Z-scale

Any data point can be converted to “Standard Units” by first subtracting the mean and then dividing by the SD.

To show how this works, consider Billy Beane’s 10.5 extra wins (on average, per season). We are all very impressed, obviously. But how impressive is this, really, in statistical terms?

Here is where standard units come in:

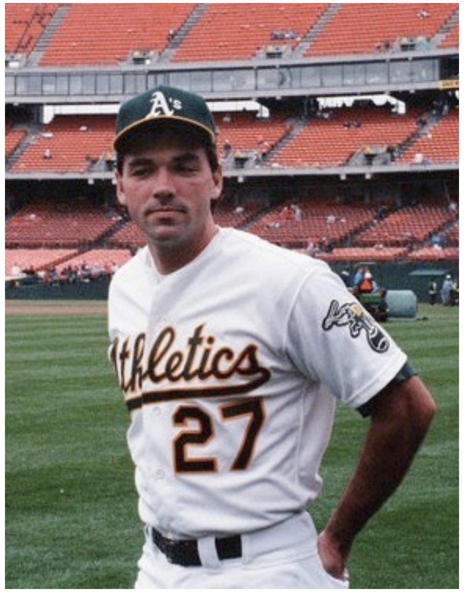
Mean = 0.319

SD = 5.359 excess wins.

Beane’s 10.5 excess wins is 10.5 - 0.319 = 10.2 wins more than average.

Now 10.2/5.359 is 1.9 SD’s above average.

$$Z = \left(\frac{x_i - \bar{x}}{s}\right) = \left(\frac{10.5 - 0.319}{5.359}\right) = 1.9$$



Billy Beane

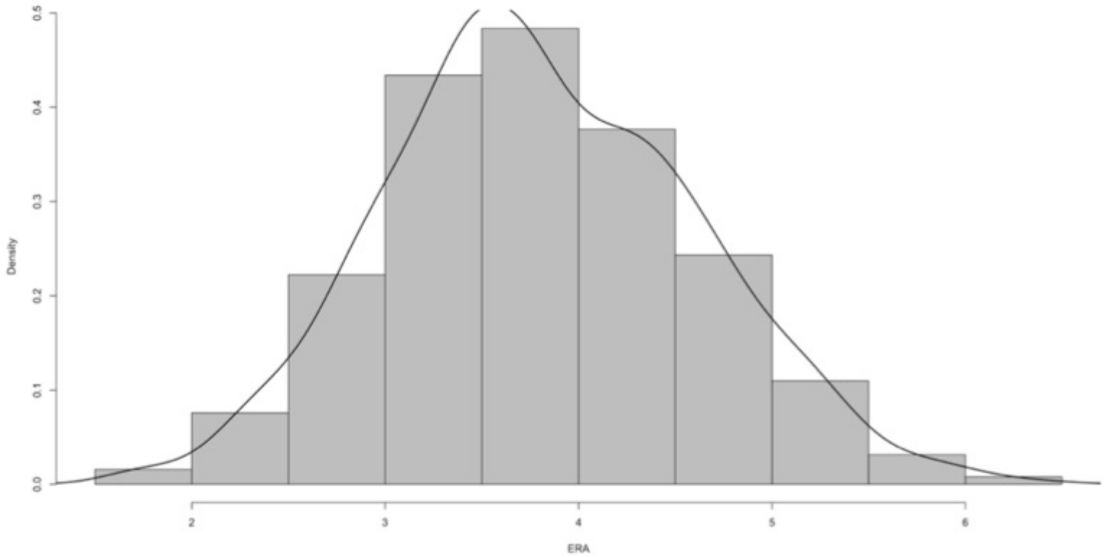
The Normal Curve applied to data

Case study 2:

Many datasets follow a "Bell Shaped" curve quite well. For these datasets the empirical rule holds precisely. In fact, every quantile can be calculated using only the mean and SD.

765 seasons for starting pitchers since 2010.

Distribution of $\boxed{\text{ERA}}$s across pitcher-seasons



Mean	3.807
SD	0.8015
N	765

You can "look-up" the frequency under a normal curve between any two points.

Source: Lahman's Baseball Database

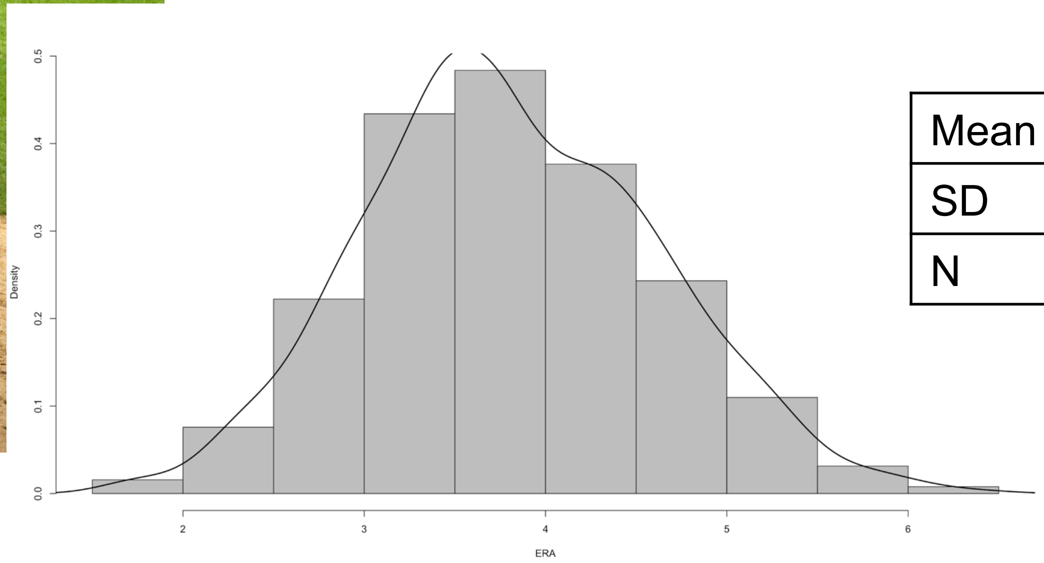
The normal curve applied to data – example pitching

So, for example, how rare has it been (in last 5 years) for a starter to have a 2.50 ERA or below?

If 2.50 was 1 SD then only 16% of pitchers would have a lower ERA.
If 2.50 was 2 SD then only 2.5% of pitchers would have a lower ERA.
2.50 is about 1.6 SDs less than the mean. It is closer to 2.5% than 16%.

Jake Arrieta, 2014
2.53 ERA

Distribution of ERAs



Mean	3.807
SD	0.8015
N	765

Normal distribution calculator

Use a calculator: http://stattrek.com/online-calculator/normal.aspx

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

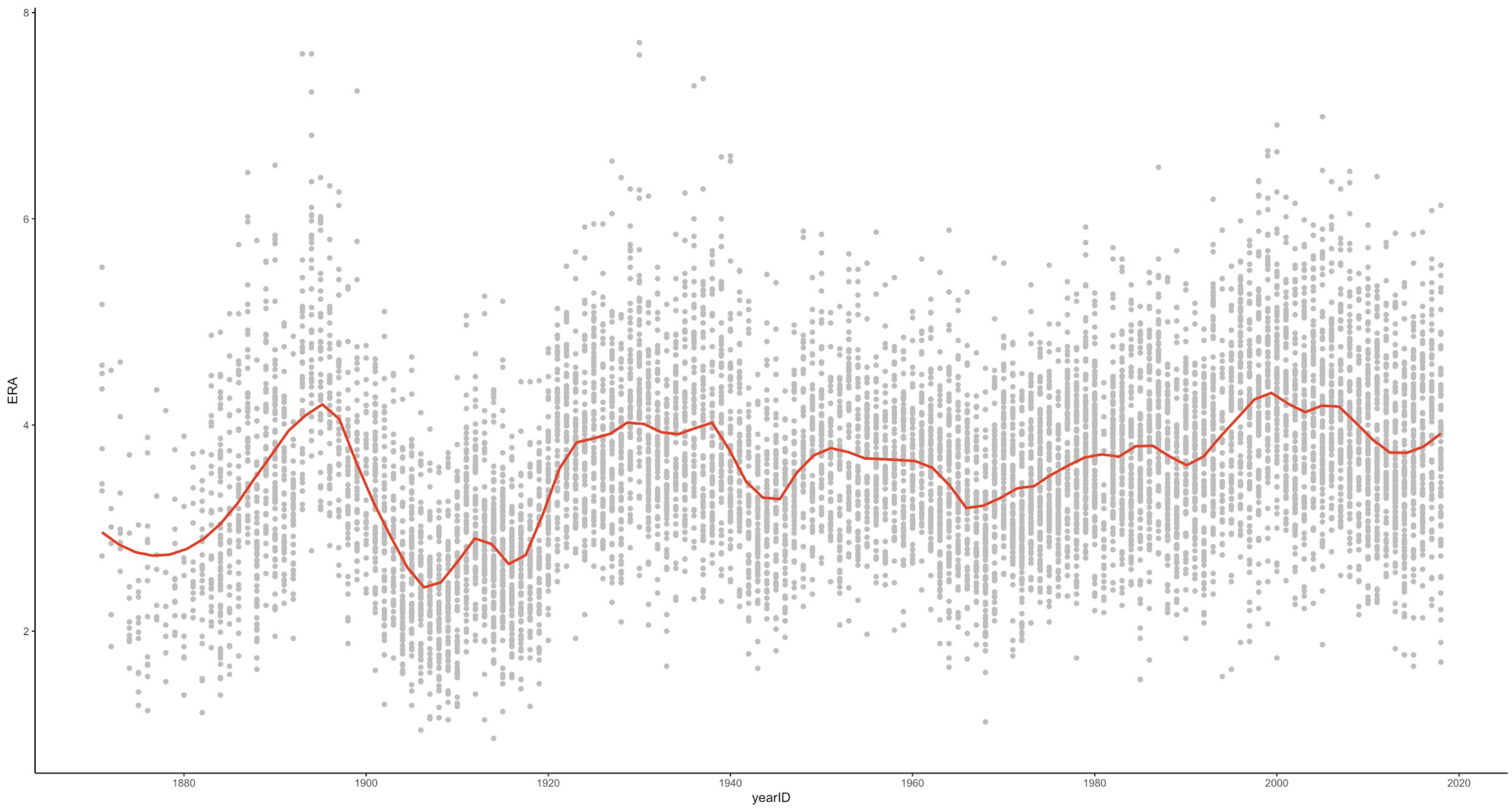
Normal random variable (x)	2.53
Cumulative probability: P(X ≤ 1.6)	0.056
Mean	3.807
Standard deviation	0.8015

$x_{\text{Arrieta, 2014}}$	2.53
Mean	3.807
SD	0.8015
N	765

You can of course use R or a calculator.

pnorm(x, mean, sd)

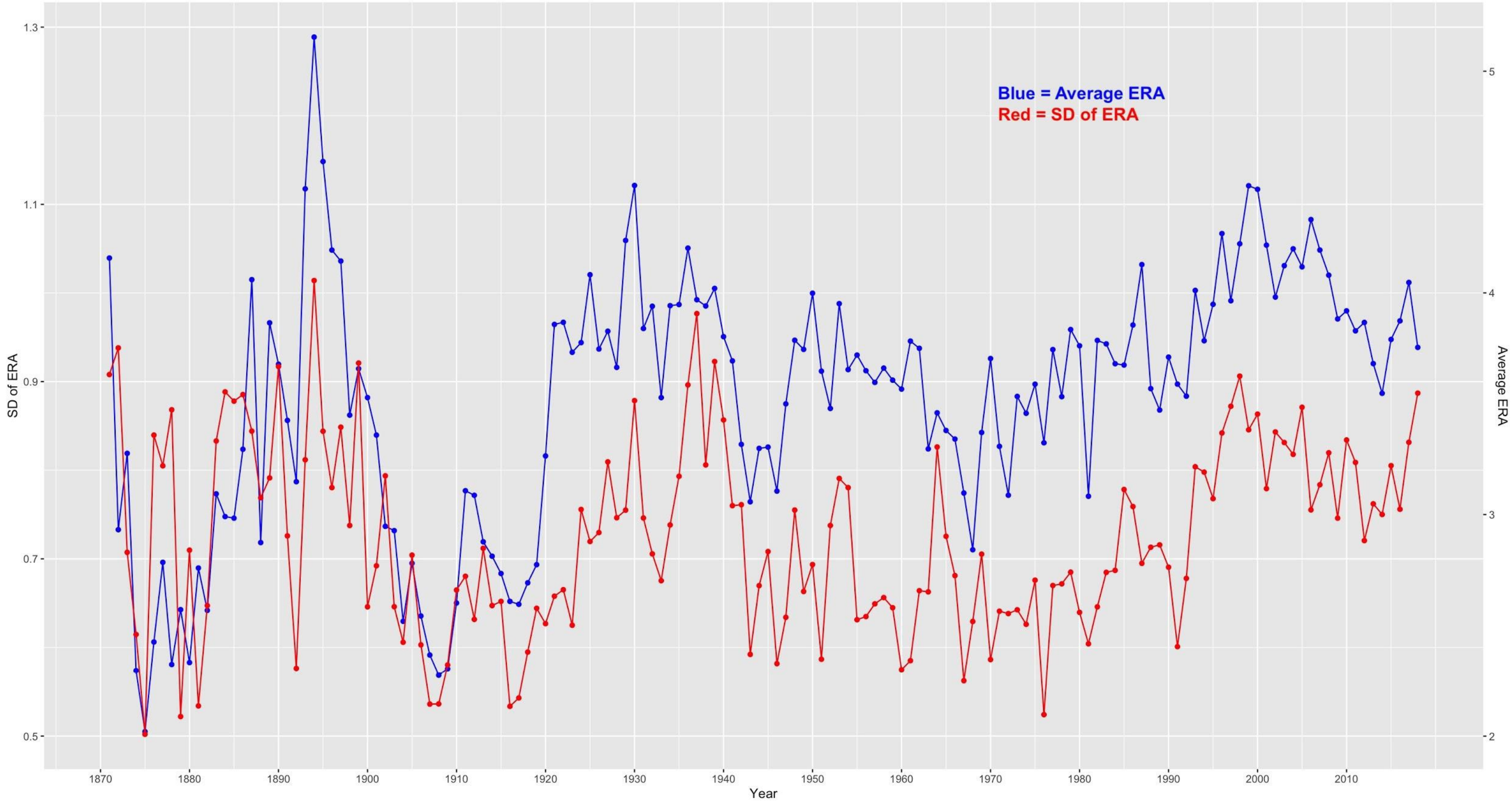
Which pitcher had the best year of all time?



Source: Lahman's Baseball Database

Which pitcher had the best year of all time?

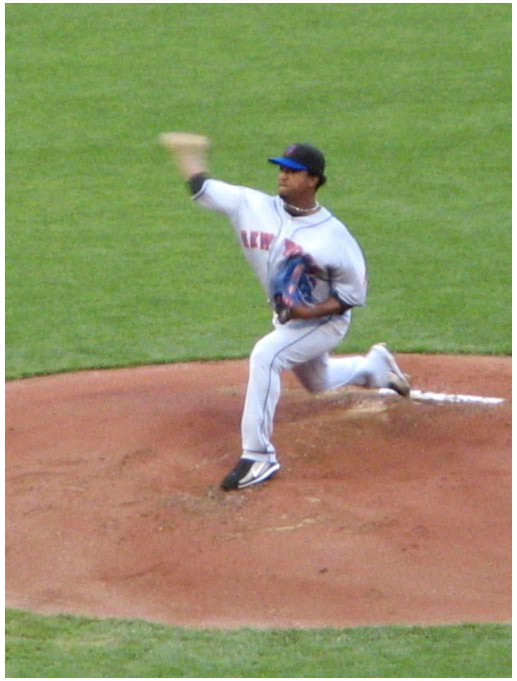
Adjust the comparison for ERA by subtracting



Source: Lahman's Baseball Database

Which pitcher had the best year of all time?

Player	Year	ERA	Standardized ERA (in SU)
Pedro Martinez	2000	1.74	-3.151
Dwight Gooden	1985	1.53	-2.998
Mark Eichorn	1986	1.72	-2.938
Greg Maddux	1994	1.56	-2.929
Greg Maddux	1995	1.63	-2.874
Dolph Leonard	1914	0.96	-2.858
Bob Gibson	1968	1.12	-2.854
Kevin Brown	1996	1.89	-2.822
Roger Clemens	2005	1.87	-2.757
Ron Guidry	1978	1.74	-2.756
Pedro Martinez	1999	2.07	-2.729
Dolf Luque	1923	1.93	-2.696
Walter Johnson	1913	1.14	-2.670
Cart Hubbel	1933	1.66	-2.599
Whitey Ford	1958	2.01	-2.583
Roger Craig	1959	2.06	-2.538
Lefty Grove	1931	2.06	-2.536



Pedro Martinez

Photo "Pedro Martinez warms up" by tingley

How about WAR as a measure of best season ?

Year	Pitcher	Team	GWAR	Z-score
1966	Sandy Koufax	LAN	11.543	4.298
1968	Bob Gibson	SLN	11.045	4.032
1985	Dwight Gooden	NYN	11.039	4.029
1997	Roger Clemens	TOR	10.97	3.993
1972	Steve Carlton	PHI	10.712	3.855
1953	Robin Roberts	PHI	10.429	3.705
1963	Sandy Koufax	LAN	10.405	3.692
1978	Ron Guidry	NYA	10.332	3.653
2000	Pedro Martinez	BOS	10.294	3.633
1972	Gaylord Perry	CLE	9.997	3.474
1964	Dean Chance	LAA	9.782	3.360
1971	Wilbur Wood	CHA	9.733	3.334
1971	Tom Seaver	NYN	9.67	3.300
1971	Vida Blue	OAK	9.67	3.300
1965	Sandy Koufax	LAN	9.595	3.260-

Why do you think modern pitchers are not appearing on this list?

(reminder: WAR is era [redacted], league and park adjusted)